Scheduling Techniques for Improving Call Capacity for VoIP Traffic in MIMO-OFDMA Networks

Marios Nicolaou, Angela Doufexi, Simon Armour and Yong Sun (Toshiba Research Europe Limited)
Department of Electrical & Electronic Engineering,
University of Bristol, Bristol, United Kingdom,
M.Nicolaou@bristol.ac.uk

***Abstract*-Fourth Generation Networks will almost invariably adopt OFDMA (Orthogonal Frequency Division Multiple Access) and MIMO (Multiple Input-Multiple Output technologies, in order to meet high data rate and Quality of Service (QoS) requirements. The Worldwide Interoperability for Microwave Access (WiMAX) MAC Layer, which is based on the IEEE 802.16 standard, is designed to support a variety of applications, including voice and multimedia services. The problem of providing QoS in broadband wireless systems is one of managing radio resources effectively. Efficient scheduling algorithms that balance the QoS requirements of each application and user with the available radio resources need to be developed. This paper considers a number of scheduling policies concentrating on real-time Voice over IP (VoIP) traffic. Via numerical results we show that the conventional notion of fairness fails to guarantee service for low latency applications such as VoIP for an increasing traffic load. We show better QoS is achieved via a greedy scheduling approach that manages to serve packets faster.**

I. INDRODUCTION

Recent years have seen the emergence of new wireless communications systems that include a variety of different applications (e.g. real time video, voice, data), radically changing the scope and design principles of future wireless systems. The IEEE 802.16 (WiMAX) [1] standard and the Third Generation Partnership Project Long Term Evolution (3GPP-LTE) [2] have been designed to address these issues. WiMAX provides wireless broadband access in metropolitan areas as an alternative to traditional wireline technologies. LTE is an extension to the UMTS mobile phone standard.

The combination of MIMO antenna technologies and Orthogonal Frequency Division Multiplexing (OFDM) has gathered significant attention over the years as a potential candidate for future wireless communication systems. OFDM transmission is particularly robust to multipath delay spread in radio channels, making it a key technology for Non Line-of-Sight (NLOS) transmission. Additional flexibility can be achieved in terms of multiple access by adaptively assigning different frequency subcarriers to different users (OFDMA). Efficient multiuser diversity (MUD) [3] can be realised by exploiting the spectral fading nature of the wireless channel, e.g. in [4, 5].

The integration of multiple transmit and/or receive antennas to OFDM can extract additional benefits in terms of data rate and link reliability, arising from additional degrees of freedom. MIMO systems are particularly attractive since they provide these performance benefits without any requirement of extra bandwidth. Due to these obvious benefits, both WiMAX and 3GPP have assumed that the downlink of the air interface would be OFDM/OFDMA based, with MIMO support.

Support for QoS is a fundamental part of the WiMAX MAC-layer design. QoS control is maintained by a connection-oriented MAC architecture, where all downlink and uplink connections are controlled by the serving Base Station (BS). To support a wide variety of applications, WiMAX defines five scheduling services: Unsolicited grant services (UGS), Real-time polling services (rtPS), Non-real-time polling service (nrtPS), Best-effort (BE) service and Extended real-time variable rate (ERT-VR) service. This paper is concerned with VoIP traffic, which is supported by UGS and ERT-VR. UGS uses a fixed amount of bandwidth for the duration of the call, with no periodic bandwidth request or polling service and is designed to support real-time service flows that generate fixed-size data packets on a periodic basis, such as VoIP without silence suppression. ERT-VR [6] is particularly useful when voice activity detection (VAD) is implemented, resulting in a variable data rate VoIP traffic [7].

The objective of the paper is to examine a VoIP traffic system under different scheduling algorithms that achieve varying degrees of throughput and fairness in a multiuser scenario and investigate the maximum VoIP customer admittance capability of these algorithms for a specified bandwidth. Real Time-VoIP traffic imposes strict packet delay constraints. Users experiencing excessive packet timeouts are assumed to receive inadequate service. A VoIP service provider should operate within a specified service outage probability.

The remainder of the paper is organised as follows: Section II introduces the channel model, VoIP traffic characteristics and the packet scheduler structure. Packet scheduling algorithms are introduced in Section III. Section IV presents and discusses simulation results. Finally, the paper concludes with Section V.

II. SYSTEM MODEL

A. Channel and Simulation Parameters

In this study, an OFDMA system with 10MHz total bandwidth (BW) is assumed. For simulation purposes, a dedicated band of 175 KHz, split into 16 non-contiguous frequency subcarriers, is allocated exclusively for servicing VoIP traffic. It has to be noted that the relatively small bandwidth has been solely adopted for simulation efficiency purposes. Results have been verified to scale according to BW. The channel model used in the simulations is based on the spatial channel model (SCM) [8]. This model was

978-1-4244-2515-0/09/$25.00 ©2009 IEEE

developed by ETSI 3GPP-3GPP2 to help standardise the outdoor evaluation of mobile systems. In accordance to [9], an urban micro 3GPP tapped delay line (TDL) channel model is generated. Table 1 summarizes key channel parameters used in this paper. A Single User-MIMO (SU-MIMO) approach, where both spatial layers of the MIMO channel are assigned to the same user is assumed. Non-contiguous frequency blocks are used to ensure frequency diversity. A long term average SNR=0dB is assumed for all users.

Table 1: Channel and Simulation Parameters

Parameter	Value
FFT size	1024
Useful Subcarriers	768
Guard Interval Length	176
Subcarrier Frequency Spacing	10.94 KHz
Symbol Duration	102.9 μS
MAC Frame Duration	5 ms
No. Rx Antennas	2
No. Tx Antennas	2
Exponentially Weighted Window (t_c) value	1000

B. VoIP Traffic Characteristics

The ERT-VR VoIP transmission with VAD can be modelled as a variable bit rate (VBR) model with active and silent periods representing characteristics of VBR traffic. A two state Markov process is used to model VoIP as indicated in Figure 1. The alternating periods of activity and silence are exponentially distributed with average durations 1/μ and 1/λ respectively. Therefore the total fraction of time the voice source is active is λ/(μ+λ). Each VoIP session is either in the Active or Inactive State. The duration at each state is exponentially distributed. During the active state, fixed sized packets of 32 bytes are generated at a constant interval of 20ms.

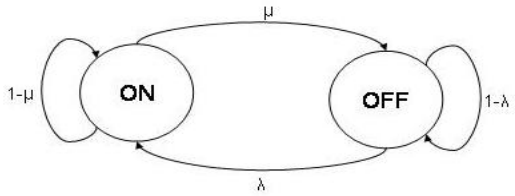

Figure 1: Two-state Markov chain for VoIP period transitions

Table 2 summarises the packet generation parameters and arrival/service distributions for VoIP traffic. The maximum tolerable packet delay for VoIP is set to 30ms. Acceptable VoIP service is maintained for a packet timeout ratio less than 4% [10]. Users exceeding this ratio are in QoS outage. Service Providers should aim to constrain QoS outage to a minimum.

Table 2: VoIP traffic distribution parameters

Component	Distribution	Parameters
Active state duration (ON)	Exponential	Mean =0.4 s
Inactive state duration (OFF)	Exponential	Mean =0.6s
Packet Inter-arrival rate within a burst	Fixed	r = 5packet/s
Probability of transition from active to inactive state	N/A	μ=0.6
Probability of transition from active to inactive state	N/A	λ=.4

C. Packet Scheduler Structure

The packet scheduler operating at the MAC layer is the critical element delivering QoS. Figure 2 shows a generalised packet scheduler structure. The packet scheduling system at the BS consists of three blocks: a packet classifier (PC), a buffer management block (BMB), and a packet scheduler (PS). The packet classifier classifies incoming packets according to their types and QoS profiles, and forwards them to buffers in BMB. The BMB maintains QoS statistics such as the arrival time and the delay deadline of each packet, the number of packets, and the head-of-line (HOL) delay in each buffer. Finally, the PS transmits packets to users according to the scheduling priority.

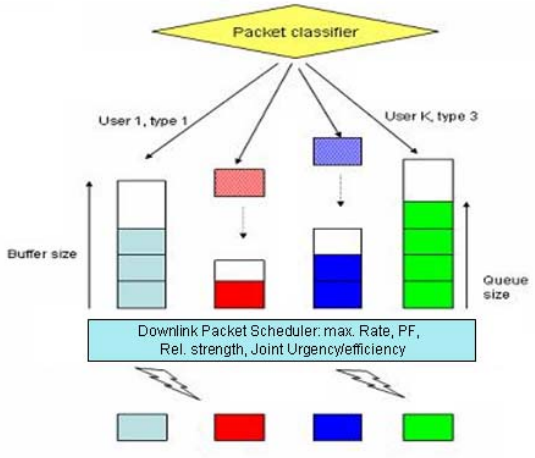

Figure 2: Generalised Packet scheduler structure

III. PACKET SCHEDULING ALGORITHMS

A. Maximum Rate Scheduling

Maximum rate scheduling has long been recognized as an effective method of enhancing the throughput of wireless channels, albeit at the expense of compromising fairness amongst users experiencing different fading levels. Assuming K users request service from a BS, each having an estimated supportable rate $R_k(t)$, at the t-th time transmission interval (TTI), the max. rate scheduling assigns resources to a user in a greedy fashion, according to the following criterion:

$$i = \arg\max_{1 \le k \le K} R_k(t) \qquad (1)$$

The integration of the maximum rate scheduling to OFDMA allows further exploitation of MUD in the spectral domain, consequently giving rise to increased rates, as well as improved resource allocation fairness [11]. For OFDMA, each user is required to feed back the estimated supportable rate for each frequency resource block, q, $R_k(q,t)$. The BS assigns different frequency resources of the same OFDM symbol to different users according to:

$$i = \arg\max_{1 \le k \le K} R_k(q,t) \qquad (2)$$

B. Proportional Fair Scheduling

The Maximum rate scheduling approach schedules packets according to the instantaneous channel conditions. It is possible therefore, that users that are not near the BS, having weaker channel conditions therefore, would never be able to get access to the resources required to serve their queued packets. The Proportional Fair (PF) scheduling

algorithm [3] schedules users, by keeping track of their previous utilisation $T_k(t)$, over an exponentially weighted window of length t_c. By assigning the user with the highest ratio of the current requested rate over the previous utilisation, each user is scheduled when its channel is relatively good and while the scheduling algorithm becomes perfectly fair in the long term. The BS assigns frequency resources to different users according to:

$$i = \arg\max_{1\le k\le K} \frac{R_k(q,t)}{T_k(t)} \qquad (3)$$

The PF scheduling algorithm has originally been designed for continuous, full buffer traffic. However, VoIP traffic is characterised by asynchronous packet arrival times from different users, as indicated in Figure 3 and therefore results in variable traffic demands at different instants. A phenomenon whereby users with new sessions are excessively prioritised over users with existing VoIP connections may occur. This is due to the fact that the latter group of users has already accumulated a certain utilisation $T_k(t)$ over the previous duration of the VoIP session. New customers that have not accumulated any utilisation yet are seen as under-performing by the scheduler and thus are allocated the bulk of the available resources. A deprivation of resources to existing customers can result in excessive packet timeouts and consequently in unacceptable VoIP service.

In order to alleviate this potential problem, we propose a modified version of PF scheduling, wherein new packet arrivals at the buffer are considered as independent events, for which utilisation over past transmitted packets is ignored. Additionally, instead of resetting the utilisation $T_k(t)$ of new packets to zero, a finite value is assumed, that ensures fairness for asynchronous packet arrival times as well as for the entire VoIP session.

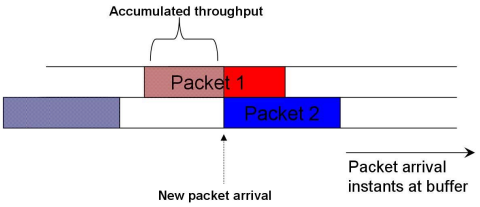

Figure 3: VoIP packet arrival model

Integration of the PF scheduling algorithm to an OFDMA scenario has been considered in [12], whereby the implications of updating the utilisation on a subcarrier basis have been discussed. Increased fairness due to the more rapid update of the fairness metric has been observed.

C. Relative Strength Scheduling

As mentioned in the previous section, PF scheduling is designed to achieve long term throughput fairness. However, the time required for the PF scheduling to settle is dependent on the weighted window length and the inherent channel variations for different users. This convergence delay can affect delay sensitive real-time applications such as VoIP.

In this section we propose an OFDMA-specific scheduling algorithm that achieved short term resource allocation fairness. By giving enhanced scheduling priority of weak users on their strong clusters, a more equally distributed resource allocation across an OFDM symbol results in a short term resource allocation fairness. The relative strength metric compares the instantaneous cluster strength of each user with the average OFDM symbol strength. Clusters that are found to be above the average symbol strength are given increased priority and clusters weaker than the average are reduced in priority. Users are selected according to the criterion:

$$i = \arg\max_{1\le k\le K} \left(\frac{|h_k(q,t)|^2}{|\overline{h_k}|^2} \right)^{\gamma} |h_k(q,t)|^2 \qquad (4)$$

The first part of the selection metric involves the cluster strength, $|h_k(q,t)|^2$, relative to the average symbol strength $|\overline{h_k}|^2$. The second part of the metric is the multiuser diversity factor. The γ parameter tunes the dependency of the metric to the relative strength parameter. Note that for γ=0, the algorithm reduces to maximum rate scheduling.

To illustrate the operation of the relative strength scheduling, Figure 4 shows the channel strength of 3 different users with highly diverse channel strengths. If maximum rate scheduling were to be employed, User 3 would not acquire any resources. With the relative strength metric (γ=2 in this case), User 3 acquires more resources over the period of an OFDM symbol, especially at subcarriers that are near their peaks. For larger channel strength variations, a higher γ value can be used, for increased dependency on the relative strength component, rather than the multiuser diversity component.

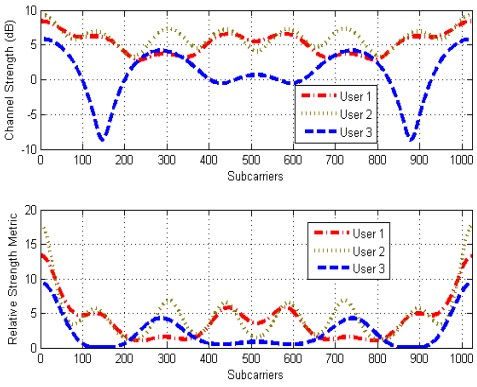

Figure 4: Relative Strength Scheduling Metric

D. Urgency Based Scheduling

So far, scheduling algorithms presented in this paper have not considered any QoS criteria for resource allocation. A joint urgency/ efficiency scheduling algorithm has been introduced in [13], that incorporates additional QoS parameters to the user selection metric, designed to serve real time (RT) and non-real time (NRT) traffic simultaneously. Two characteristics are identified for tasks associated with time constraints, the task execution time and the deadline. These two characteristics are represented by a value function in time. Two scheduling factors, the urgency of the scheduling and the efficiency of the radio resource are used to determine priority for each user. A time utility function (TUF) is used to indicate the urgency of the scheduling.

Efficiency in wireless communications is related to the usage of limited radio resources. Hence, the channel state of the available radio resources can be used as an efficiency indicator. A number of different efficiency indicators include: the current supportable date rate $R_k(q,t)$, average channel

rate $\overline{R_k(t)}$ or the ratio of the current supportable rate to the average $R_k(q,t)/\overline{R_k(t)}$. This study adopts the current supportable rate as the efficiency metric of choice.

Concentrating on RT traffic, the scheduler needs to successfully transmit packets at any time within the maximum packet lifetime to satisfy the delay requirement, 30ms, for the investigated VoIP scenario. The time utility function (TUF) for real time traffic illustrated in Figure 5 as a function of the Head of Line (HOL) packet delay is expressed as follows:

$$U_{RT}(t) = \frac{e^{-\alpha(t-c)}}{\left(1+e^{-\alpha(t-c)}\right)} \quad (5)$$

α being the slope parameter and c the location parameter of the inflection point.

The packet urgency function is given by the derivative of the TUF, U'_{RT}. What can be observed from Figure 5 is that the packet urgency starts to diminish, after the inflection point and drops to zero as the packet lifetime moves closer to the maximum tolerable packet delay. Hence, the point of inflection can be viewed as the point where the scheduler begins to "give up" on the packet, and prioritises packets that have a more realistic chance of successful transmission over the remaining time frame.

The urgency/efficiency scheduling algorithm allocates frequency resources to users according to:

$$i = \arg\max_{1\le k\le K} U'_{RT}(k) R_k(q,t) \quad (6)$$

RT packets are scheduled over a marginal scheduling time interval (MSTI) over which the urgency value is assigned a non-zero value. Outside this interval, other packets, especially Non Real Time (NRT) packets can be transmitted.

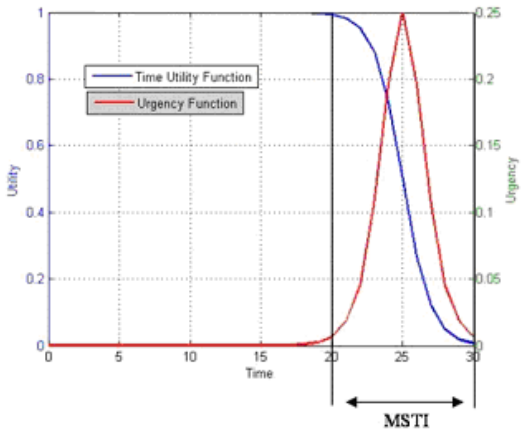

Figure 5: Utility and Urgency function for VoIP traffic (α=1, c=25)

IV. QUANTITATIVE ANALYSIS

For the assessment of the scheduling algorithms under consideration, appropriate criteria relating to VoIP requirements need to be used. For real time traffic, the primary target is the fulfilment of the packet delay requirement. Users experiencing high packet delays that lead to excessive packet timeout can result in outage. In terms of measuring overall system satisfaction, the average packet timeout ratio as well as user satisfaction as a function of the active VoIP session is required.

Figure 6 depicts the average system timeout ratio as a function of the number of VoIP users for the algorithms under consideration. As expected, the timeout ratio is a direct function of the number of users. In a highly congested system, the scarcity of the available resources results in increased packet delays that translate into high overall packet timeouts.

The max. rate achieves the lowest average packet timeout ratio. The relative strength metric (γ=2) under-performs due to poor MUD utilisation. The modified PF algorithm offers slightly superior performance than the conventional time domain PF scheduling but not enough to surpass the performance of the maximum rate algorithm. The joint urgency/efficiency metric that has been designed to simultaneously serve RT and NRT traffic gives a significantly high packet timeout ratio. This effect can be attributed to the reduced length of the MSTI. Restricting packets from being transmitted until they acquire a non-zero urgency function, results in a poor spectrum utilisation. Effectively restricting VoIP packets from being transmitted until a high HOL packet delay is reached may in fact result in some packets not acquiring the necessary resources over the remaining lifetime, especially in a highly congested system.

The packet timeout ratio is an indication of how well on average a system servicing VoIP traffic is performing. User satisfaction provides a measure of how well individual users are performing. User satisfaction of a VoIP session is defined by the average packet timeout ratio of each user, which is required to be under 4% of the total transmitted packets. Figure 7 compares the user satisfaction ratio as a function of the active VoIP users. What is surprising is that the algorithms designed for improved fairness either in terms of long term throughput (frequency domain PF) or resource allocation fairness (relative strength) fail to provide any benefits for RT traffic with stringent packet delay constraints. The joint urgency /efficiency algorithm that incorporates additional QoS criteria for resource allocation also seems to fail when a single traffic scenario is considered.

These results lead to the hypothesis that the conventional notion of fairness, as it has been established for traffic with no delay constraints, fails for real time traffic. In Figure 8 we consider the cumulative distributions of the average packet delays, for 60 active VoIP users. The highly fair algorithms (freq. domain PF, relative strength) result in a more uneven delay distribution with a higher mean. The joint Urgency /Efficiency algorithm results in a highly dispersed distribution with most of the packet delays concentrating in the MSTI region. It can therefore be deduced that this algorithm fails to provide benefits for a single VoIP traffic scheme, since packets suffer excessive delays that reduce the system's ability to exploit MUD. The maximum rate algorithm results in the fairest and lowest packet delay distributions as compared with the other scheduling algorithms.

This observation explains the superiority of max. rate scheduling for VoIP. Fairness oriented algorithms tend to maintain packets in the buffer for longer, due to a more equally distributed resource assignment. However, this effect results in buffer congestion, since packets remain in the buffer for longer. In contrast, a greedy scheduling approach (i.e. the max. Rate scheduler) tends to serve strong users faster, hence removing them from the buffer. Due to the small VoIP packet size (fixed at 32 bytes), transmission of a packet within the required lifetime for a user with a strong channel does not require excessive resources. Therefore, overall system fairness is maintained by prioritising strong users requiring fewer resources for successful packet transmission, consequently allowing weaker users to get access to more

resources that in turn will allow their packets to be successfully transmitted within the required time frame.

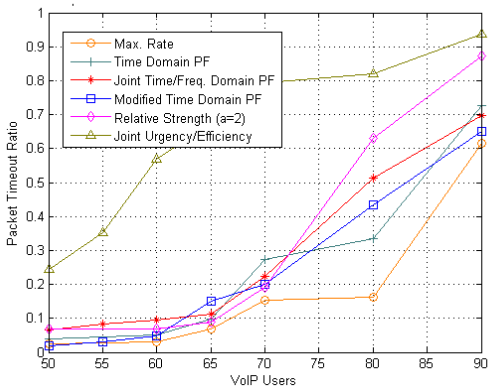

Figure 6: Packet timeout ratios for different scheduling policies

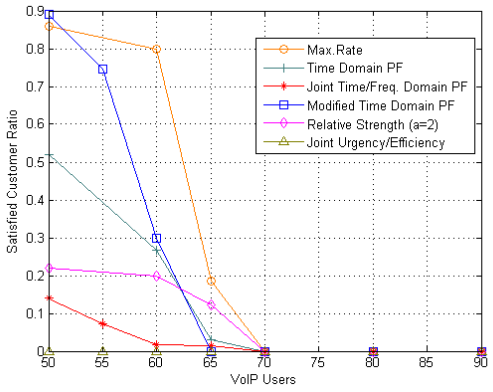

Figure 7: User satisfaction ratios for different scheduling approaches

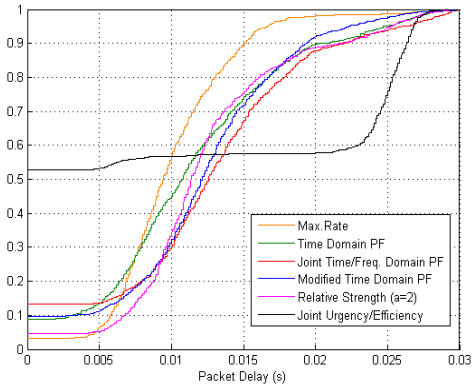

Figure 8: VoIP packet delay distributions

V. CONCLUSIONS

In this paper, several packet scheduling algorithms, incorporating varying degrees of throughput maximisation, fairness and QoS provisioning have been tested for VoIP traffic, in a MIMO-OFDMA system. Analysis has shown that traditional fairness criteria, (long term average throughput or short term resource allocation fairness) fail to efficiently serve VoIP traffic. A scheduling algorithm that incorporates head-of line packet delay as an additional QoS criterion, originally designed for serving simultaneously a mixture of traffic types has been considered, resulting, however, in excessive packet delays. An aggressive approach towards resource assignment results in an overall performance enhancement, allowing more VoIP users to be admitted in a specified system, whilst meeting their QoS requirements. Overall system fairness is achieved by prioritising strong users, freeing up resources for weaker users faster, allowing them to also meet their QoS requirements.

ACKNOWLEDGMENTS

The authors wish to acknowledge the financial support of EPSRC and Toshiba Research Europe Limited (TREL).

REFERENCES

[1] "Air Interface for fixed broadband wireless access systems," *IEEE STD 802.16-2004, October 2004*.

[2] 3GPP, *Technical Specification Group Radio Access Network; Requirements for E-UTRA and E-UTRAN (R7)'*, vol. 3GPP TR 25.913 V7.3.0, March 2006, http://3gpp.org/ftp/Specs/html-info/25913.htm

[3] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *Information Theory, IEEE Transactions on*, vol. 48, pp. 1277-1294, 2002.

[4] P. Svedman, S. K. Wilson, J. Leonard, L. J. Cimini, and B. Ottersten, "Opportunistic Beamforming and Scheduling for OFDMA Systems," IEEE Vehicular Technology Conference, VTC *Spring '04*.

[5] S. Olonbayar and H.Rohling, "Multiuser Diversity and subcarrier allocation in OFDM transmission systems," 11th Int. OFDM Workshop (InOwo), September 2006.

[6] H. Lee, T. Kwon, and D.-H. Cho, "Extended-rtPS Algorithm for VoIP Services in IEEE 802.16 systems," IEEE International Conference on Communications, ICC '06.

[7] R. V. Prasad, A. Sangwan, H. S. Jamadagni, C. M.C, R. Sah, and V. Gaurav, "Comparison of Voice Activity Detection Algorithms for VoIP " *Proceedings of the Seventh International Symposium on Computers and Communications (ISCC'02)* IEEE Computer Society, 2002 pp. 530

[8] D. S. Baum, J. Hansen, and J. Salo, "An interim channel model for beyond-3G systems: extending the 3GPP spatial channel model (SCM)," IEEE Vehicular Technology Conference, VTC '05-Spring.

[9] M. Tran, G. Zaggoulos, A. Nix, and A. Doufexi, "Mobile WiMAX: Performance Analysis and Comparison with Experimental Results," IEEE Vehicular Technology Conference, VTC '08-Fall.

[10] H. Hassan, J. M. Garcia, and C. Bockstal, "Aggregate Traffic Models for VoIP Applications," IEEE Conference on Digital Telecommunications, ICDT '06.

[11] L.-C. Wang and W.-J. Lin, "Throughput and fairness enhancement for OFDMA broadband wireless access systems using the maximum C/I scheduling," IEEE Vehicular Technology Conference VTC '04-Fall.

[12] W. Anchun, X. Liang, Z. Shidong, X. Xibin, and Y. Yan, "Dynamic resource management in the fourth generation wireless systems," IEEE International Conference on Communication Technology, ICCT 2003.

[13] S. Ryu, B. Ryu, H. Seo, and M. S. A.-M. Shin, "Urgency and Efficiency based Packet Scheduling Algorithm for OFDMA wireless system," IEEE International Conference on Communications, ICC 2005.